**SFWRTECH 4DM3**

**Data Mining**

**Comparing Classifiers**

**Course Project**

Foriben Ngang

(400281290)

**Computational Time**

|  |  |  |  |
| --- | --- | --- | --- |
| Classification | Computational Time(s) | | |
|  | Training | Testing | Total |
| K Nearest Neighbors | 0.012002229690551758 | 0.07776927947998047 | 0.089771509170532228 |
| AdaBoost | 1.1538090705871582 | 0.04035830497741699 | 1.19416737556457519 |
| Support Vector Machine | 14.268693447113037 | 0.055169105529785156 | 14.323862552642822156 |

**Cross Validation for parameter selection**

**K Nearest Neighbors**

Computational Time: 9.85985255241394s

Computed cross validation on 1-50 number of neighbors, with 10 folds.

**15** is the best k with the lowest error.

**Chart, histogram

Description automatically generated**

**Support Vector Machine**

Computation Time: 3005.5323009490967s

Computed cross validation on 1-20 regularization parameters with 3 folds.

**11** is the best parameter with the lowest error.

**Chart, line chart

Description automatically generated**

**ROC Curves**

Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generated

**Confusion Matrix**

**K Nearest Neighbors AdaBoost**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | True Positive  331 | False Positive  33 | | False Negative  91 | True Negative  4020 | | |  |  | | --- | --- | | True Positive  352 | False Positive  33 | | False Negative  70 | True Negative  4020 | |

|  |  |
| --- | --- |
| True Positive  349 | False Positive  26 |
| False Negative  73 | True Negative  4027 |

**Support Vector Machine**

|  |  |
| --- | --- |
|  |  |

**Python Code**

See Attachment

**Project Summary**

**Dataset**

In this project, HTRU2 data set was selected from the UCI Machine learning repository. This dataset describes a sample of pulsar candidates collected during the High Time Resolution Universe Survey. Pulsars are a rare type of neutron star that produces radio emissions detectable on earth. As pulsars rotate, their emission beam sweeps across the sky, and this crosses our line of sight which produces a detectable pattern of broadband radio emission. The pattern will repeat while pulsars rotate rapidly. Therefore, the pulsar search involves looking for periodic radio signals and large radio telescopes. Each Pulsar will produce different emission patterns which vary with each rotation. This dataset recorded is to automatically label pulsar candidates to facilitate rapid analysis. This dataset is a legitimate pulsar example with most negative classes and a minority of positive classes. The dataset contains 17,898 examples where 1,639 are positive and 16,259 are negative examples.

This dataset contains 9 attributes – 8 continuous variables and a single class variable. The first four variables are the statistics obtained from the integrated pulse profile (the continuous variables describe a longitude-resolved version of signal that has been averaged in both time and frequency) and the remaining four variables are obtained from the DM-SNR curve, as showed in below chart.

|  |  |
| --- | --- |
| **Attribute** | **Description** |
| 1 | Mean of the integrated profile |
| 2 | Standard deviation of integrated profile |
| 3 | Excess kurtosis of the integrated profile |
| 4 | Skewness of the integrated profile |
| 5 | Mean of the DM-SNR curve |
| 6 | Standard deviation of the DM-SNR curve |
| 7 | Excess kurtosis of the DM-SNR curve |
| 8 | Skewness of the DM-SNR curve |
| 9 | Class |

The Dataset was separated into two sets 75% (13,423) for training and 25% (4,475) for testing. All the data from the dataset were numeric attributes and for the three classifiers, 13,423 data points were used to train the algorithms and 4,475 data points were used for testing the algorithms. The accuracy of K Nearest Neighbors was 97.14% , the accuracy of AdaBoost was 97.70% and that of Vector Support Machine was 97.74%.